WELCOME TO THE 2023 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

NDACAN SUMMER TRAINING SERIES SCHEDULE 2023

- July 5 — Introduction to NDACAN and the Administrative Data Series
- July 12 — New Data Acquisition: CCOULD Data
- July 19 — Causal Inference Using Administrative Data
- July 26 — Evaluating and Dealing with Missing Data in R
- August 2 — Time Series Analysis in Stata
- August 9 — Data Visualization in R

SESSION AGENDA

- What is causal inference?
- How can we do causal inference?
- Causal inference in practice
- Goal: build general knowledge and intuition

WHAT IS CAUSAL INFERENCE?

WHAT'S THE GOAL?

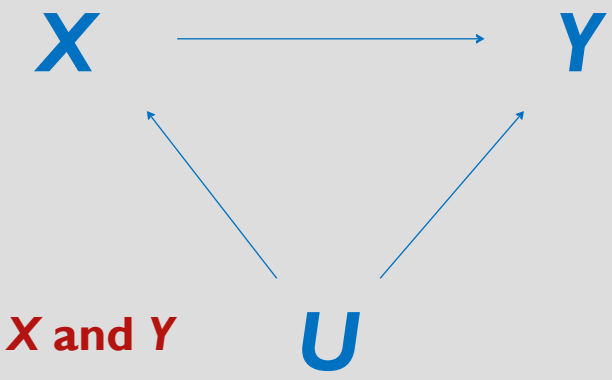
- We want to know whether *X* causes *Y*
- Can help us adjust behavior, allocate resources, inform policy, etc.

- For unit *i* the average causal effect of a given treatment = $Y_i(1) - Y_i(0)$

WHAT'S THE PROBLEM?

- We cannot observe the counterfactual—a unit (person, school, city, etc.) either does, or does not, receive the treatment
- "The fundamental problem of causal inference" and potential outcomes framework
  - A case where Wikipedia is useful: Rubin causal model
  - (https://en.wikipedia.org/wiki/Rubin_causal_model)

- We are trying to overcome various issues of *endogeneity* and *confounding*:
  1. **Selection bias**, or the fact that individuals who receive a given treatment are often very different than those who do not
  2. **Unobserved heterogeneity**, or the fact that these individuals likely differ in ways that we cannot/do not measure



U confounds* the relationship between *X* and *Y

CONTEXT

- Two aspects of causal inference:
  1. Causal Inference as an actual academic field
  2. Trying to make causal inferences about an empirical relationship

- At the end of the day, we are just trying to eliminate confounds and identify or create a control group that is as similar as possible to the treatment group

HOW CAN WE DO CAUSAL INFERENCE?

DATA

- Survey data vs. administrative data
- Surveys are when individuals are…surveyed:
  - Examples: Add Health, Fragile Families, NLSY

DATA

- Survey data vs. administrative data
- Surveys are when individuals are…surveyed:
  - Easier to access (usually)
  - Can get self-reported and other hard-to-measure data
  - But can typically only do condition-on-observables (matching, weighting, etc.) or fixed effects strategies

DATA

- Survey data vs. administrative data
- Administrative records are officially collected/recorded by organizations—in social science, that's usually government agencies
  - Examples: National Child Abuse and Neglect Data System (NCANDS) or Adoption and Foster Care Analysis and Reporting System (AFCARS). School and crime records (usually at city or state level) also very common.

DATA

- Survey data vs. administrative data
- Administrative records:
  - Often harder to access
  - Will not have some important measures *and* doesn't capture some of the population
    - "Who is missing from administrative data", Georgetown University:
    - https://mccourt.georgetown.edu/news/who-is-missing-from-administrative-data/
  - But typically has the level of detail, timing, and sufficient sample size to (potentially!) utilize conventional causal inference techniques
    - And are officially recorded (i.e., we actually observe a child's test score, instead of asking them to remember and report what their score was)

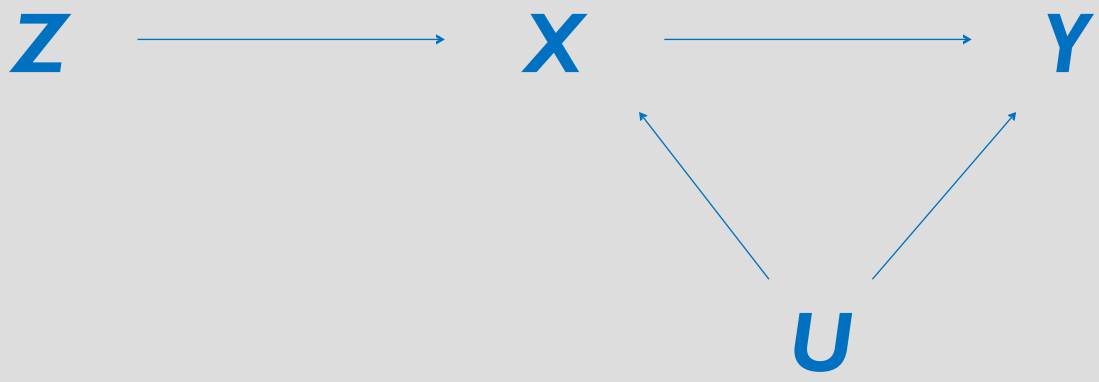
RANDOMIZED CONTROLLED TRIAL (RCT)

- The "gold standard"
- Randomly assign some people to the treatment and others to the control
- But in the social sciences (especially when thinking about the child welfare system) we very often cannot practically or ethically conduct an RCT

NATURAL EXPERIMENT OR QUASI-EXPERIMENTAL

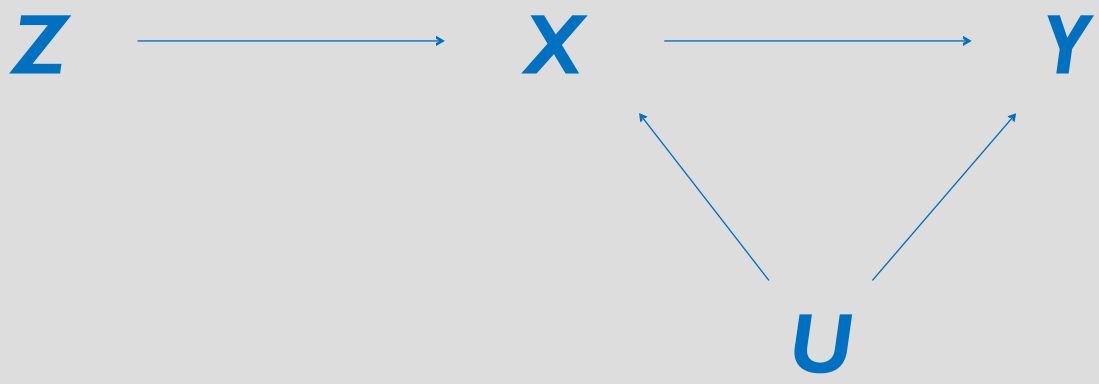
- There is some randomization that has occurred—whether by "nature" or some type of statistical procedure
- Some debate about differences between natural experiment vs. quasi-experimental, but for now we will use them synonymously to distinguish them from an RCT
- Now we'll dive into a few of the most common approaches

- Identifying and utilizing a third variable (referred to as an *instrument*) that can affect the outcome *only through its effect on the predictor*
  - Z only influencing Y through X is the **exclusion restriction**



INSTRUMENTAL VARIABLES (IV)

- The intuition here is the instrument should only cause changes in the treatment, and therefore can be used to recover a treatment effect

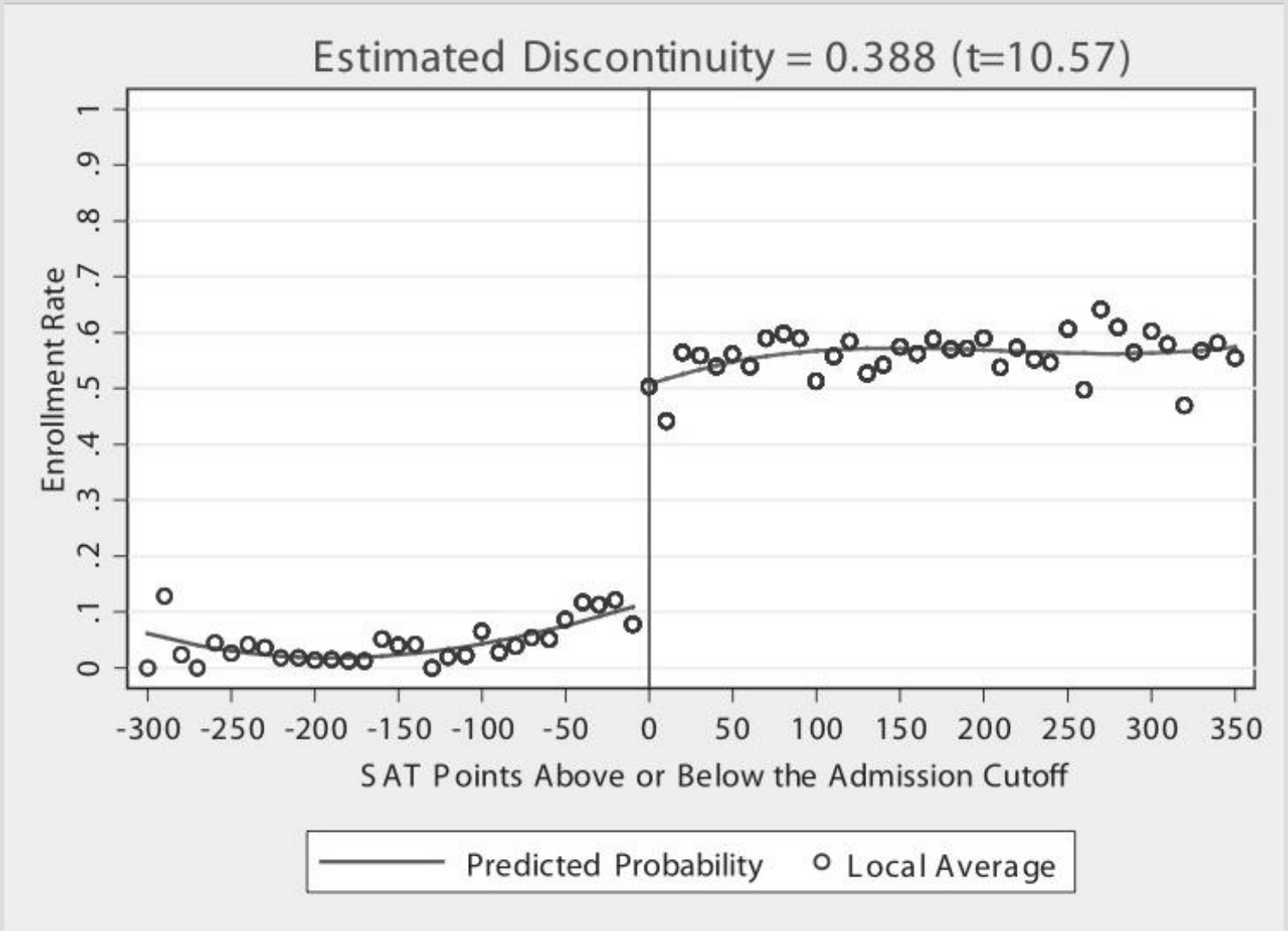


INSTRUMENTAL VARIABLES (IV)

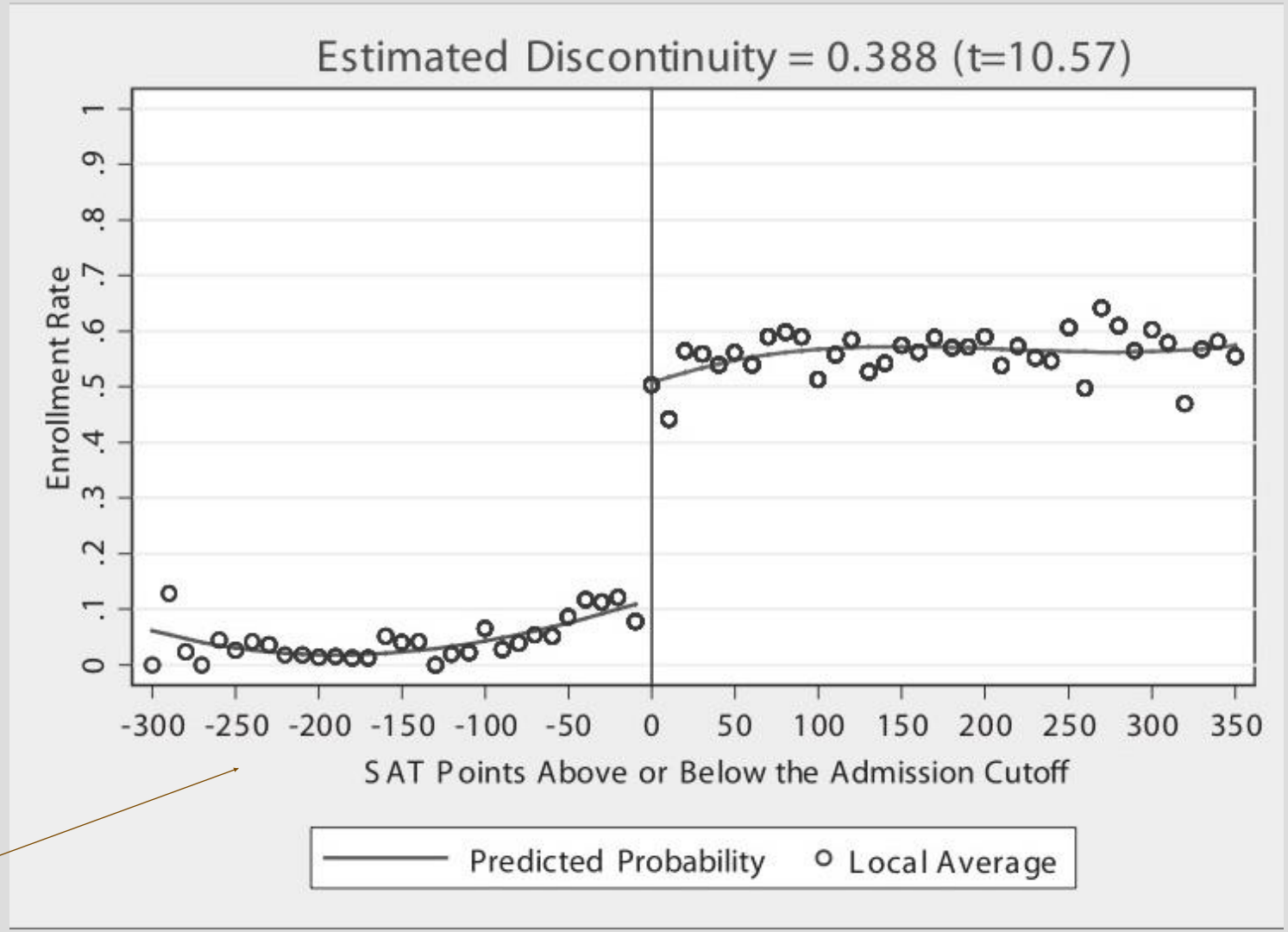
- IV requires a lot of big assumptions
- Recently there have been growing concerns about replicability, "weak" instruments, and IV in general
  - Further reading:
    - Andrews, Stock, and Sun (2018)
    - Mellon (2020)
    - Bound, Jaeger, and Baker (1995)
    - Lal et al (2021)

REGRESSION DISCONTINUITY (RDD)

- Leverages a *cutoff score* that exists in the real world and is used to sort/treat/select etc. individuals into (and out of) something
- In the real world, anyone above a certain score receives X, anyone below the score receives nothing (or some alternative)
- We can compare individuals just *on either side of the cutoff*, who should be essentially identical (on average)

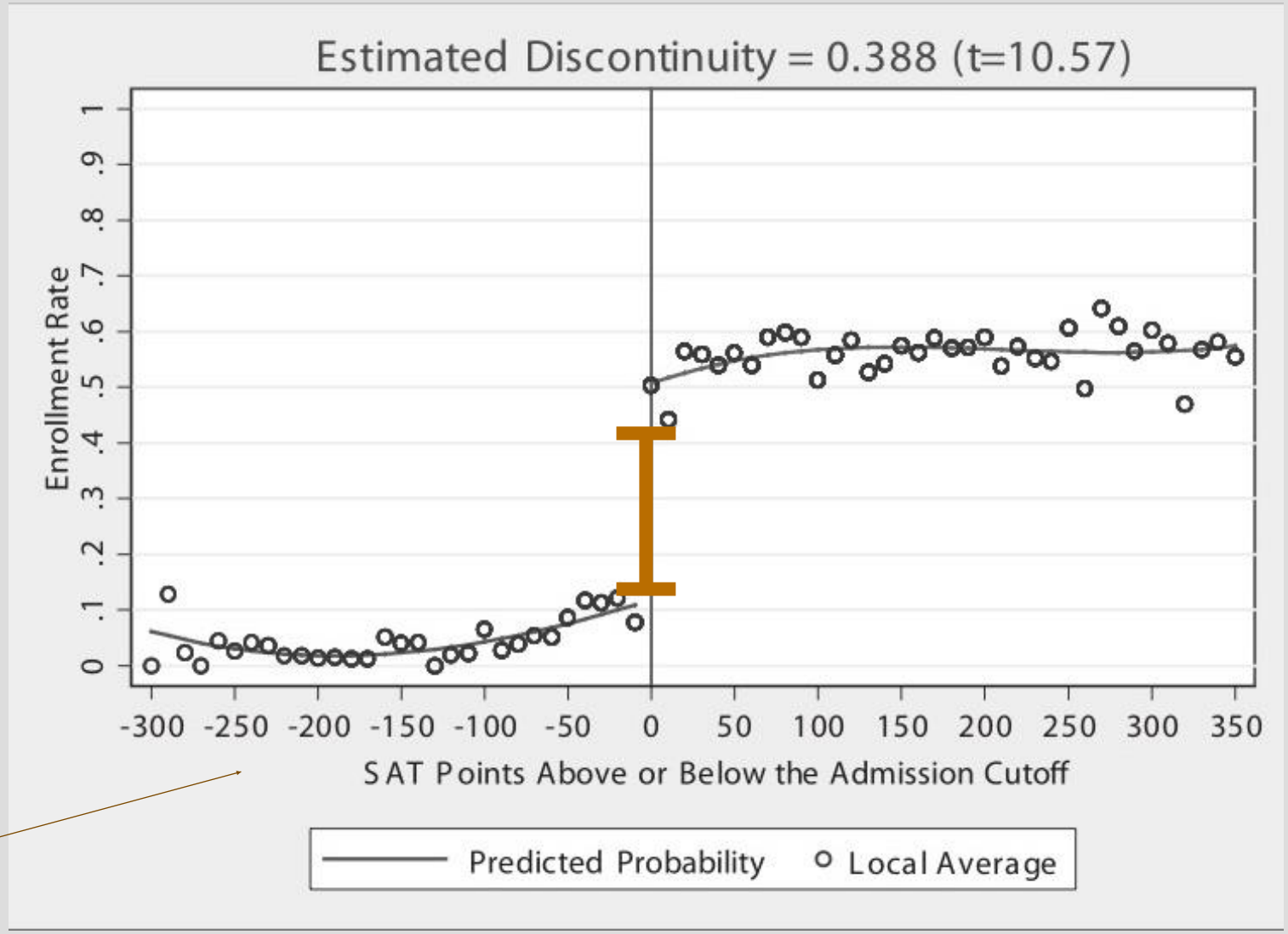


Source: Cunningham, Causal Inference The Mixtape, Ch. 6. (2000).



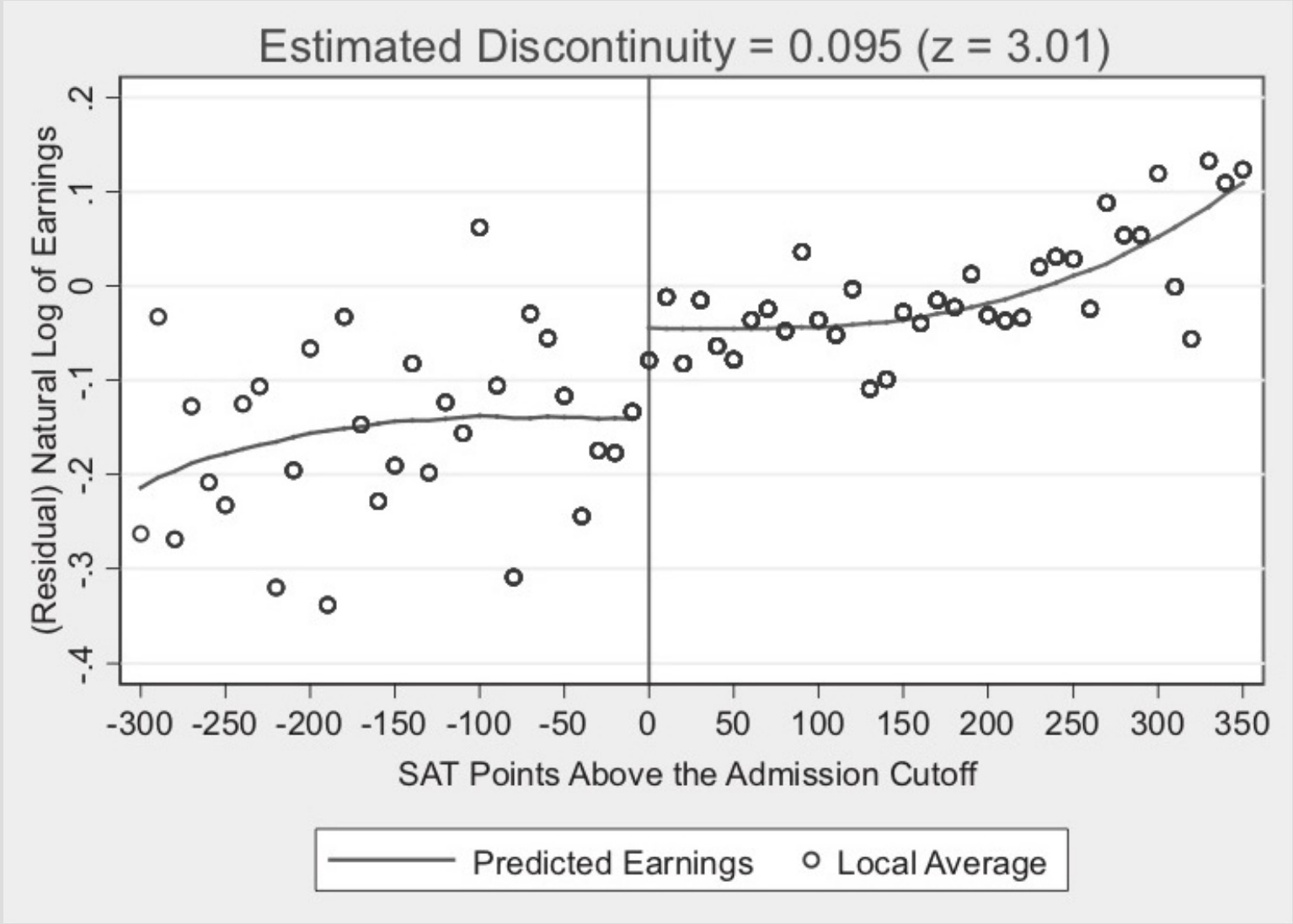
"Running variable"

Source: Cunningham, Causal Inference The Mixtape, Ch. 6. (2000).



"Running variable"

Source: Cunningham, Causal Inference The Mixtape, Ch. 6. (2000).

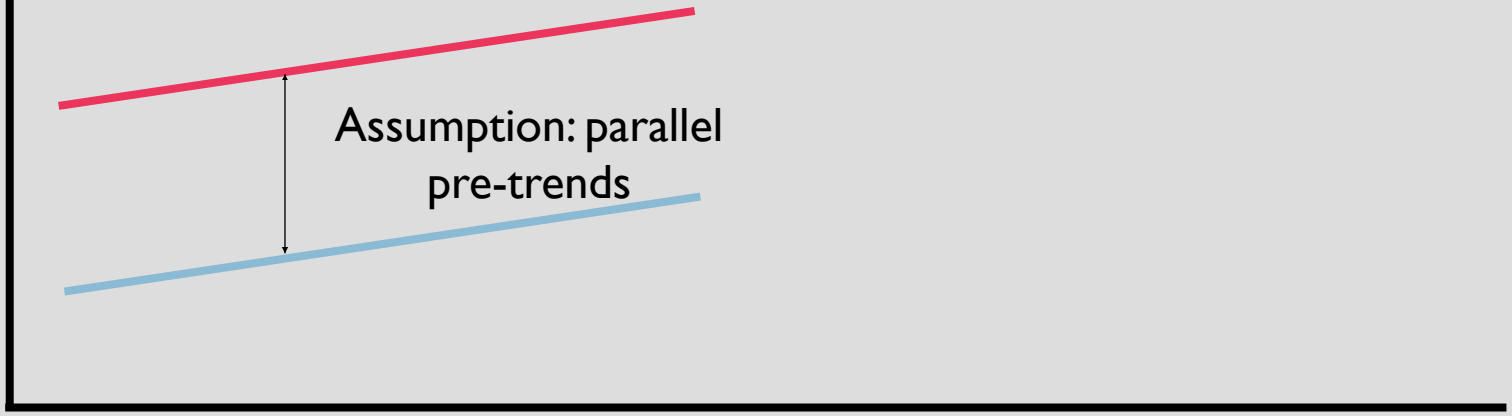


Source: Cunningham, Causal Inference The Mixtape, Ch. 6. (2000).

REGRESSION DISCONTINUITY (RDD)

- Increasingly popular in practice because these cutoffs are used in the real world all the time
- Also relies on less precarious assumptions (e.g., error term does not jump at cutoff) and is relatively clear conceptually
- Fuzzy RDD vs. Sharp RDD

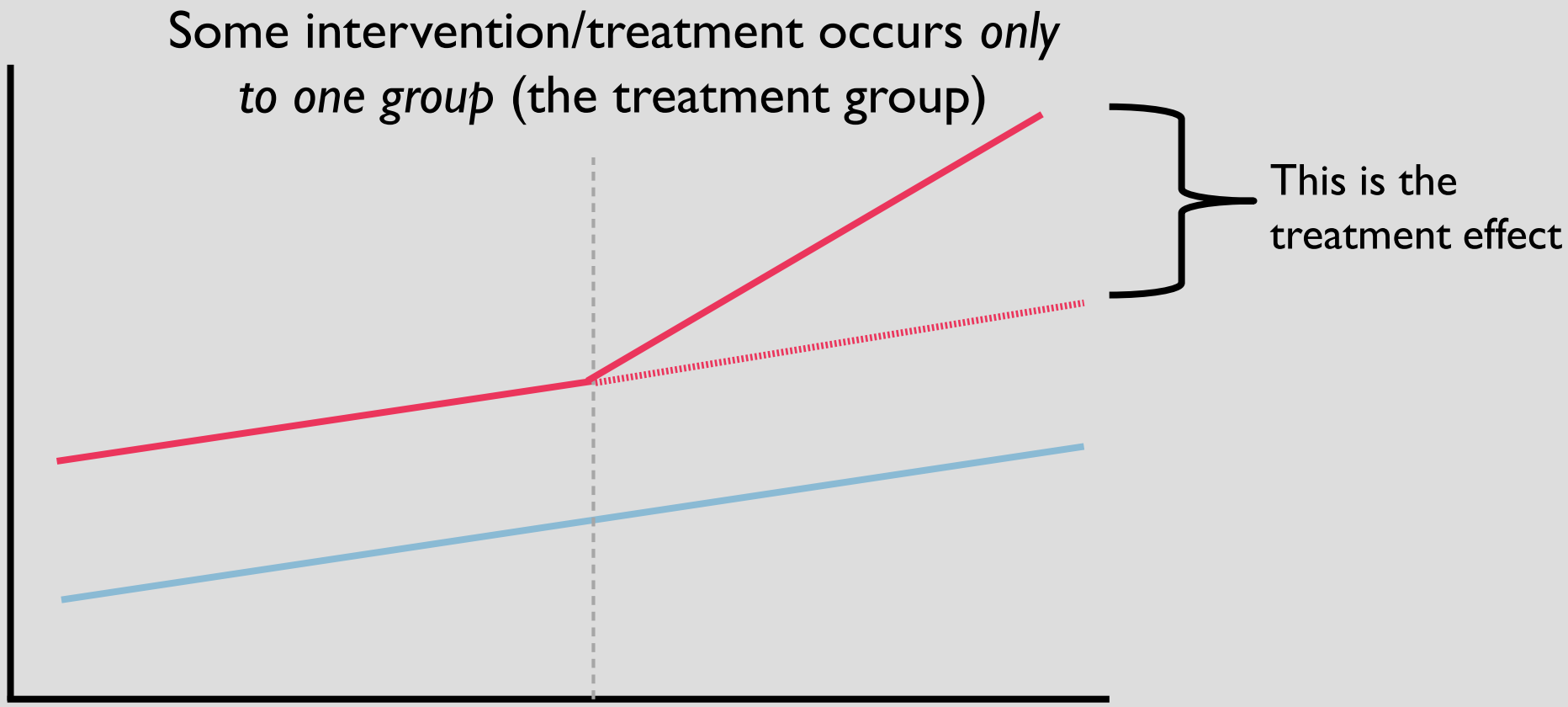
Some intervention/treatment occurs *only to one group* (the treatment group)



DIFFERENCE-IN-DIFFERENCES (DID)

1. Before-after for treatment (B-A)
2. Before-after for control (C-D)
3. Calculate difference between the difference for 1 and the difference for 2 (hence difference in differences!)…

$Y = (B\text{-}A) - (C\text{-}D)$



DIFFERENCE-IN-DIFFERENCES (DID)

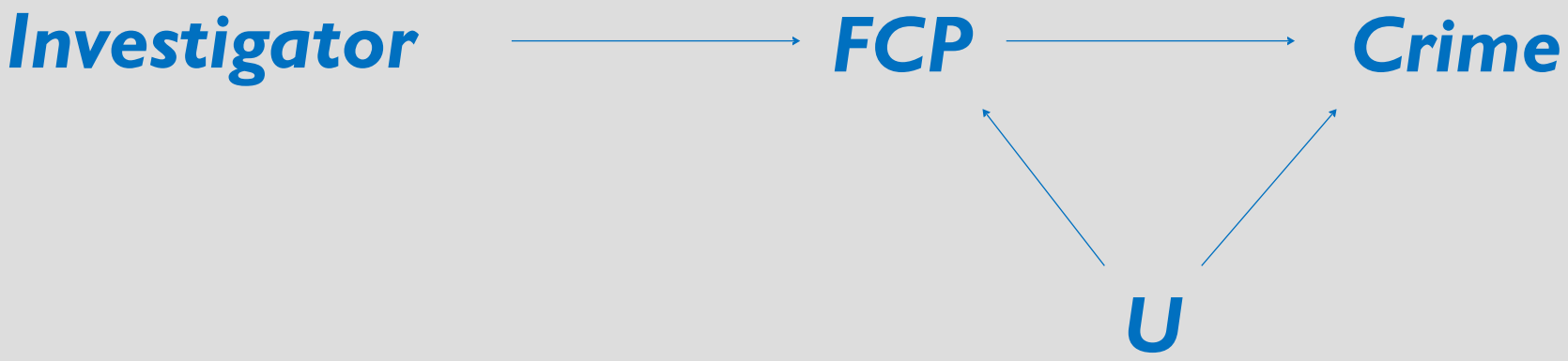
- There are lots of complications with DiD, especially when there are multiple treated units that are treated at multiple times
  - Two-way fixed effects (TWFE) estimator gets very tricky very quickly
  - Goodman-Bacon (2021) and Callaway and Sant'Anna (2021)
    - Two resources to provide additional description and context:
      - https://causalinf.substack.com/p/callaway-and-santanna-dd-estimator
      - http://resources.oliviajhealy.com/TWFE_Healy.pdf

CAUSAL INFERENCE IN PRACTICE

BARON AND GROSS (2022)

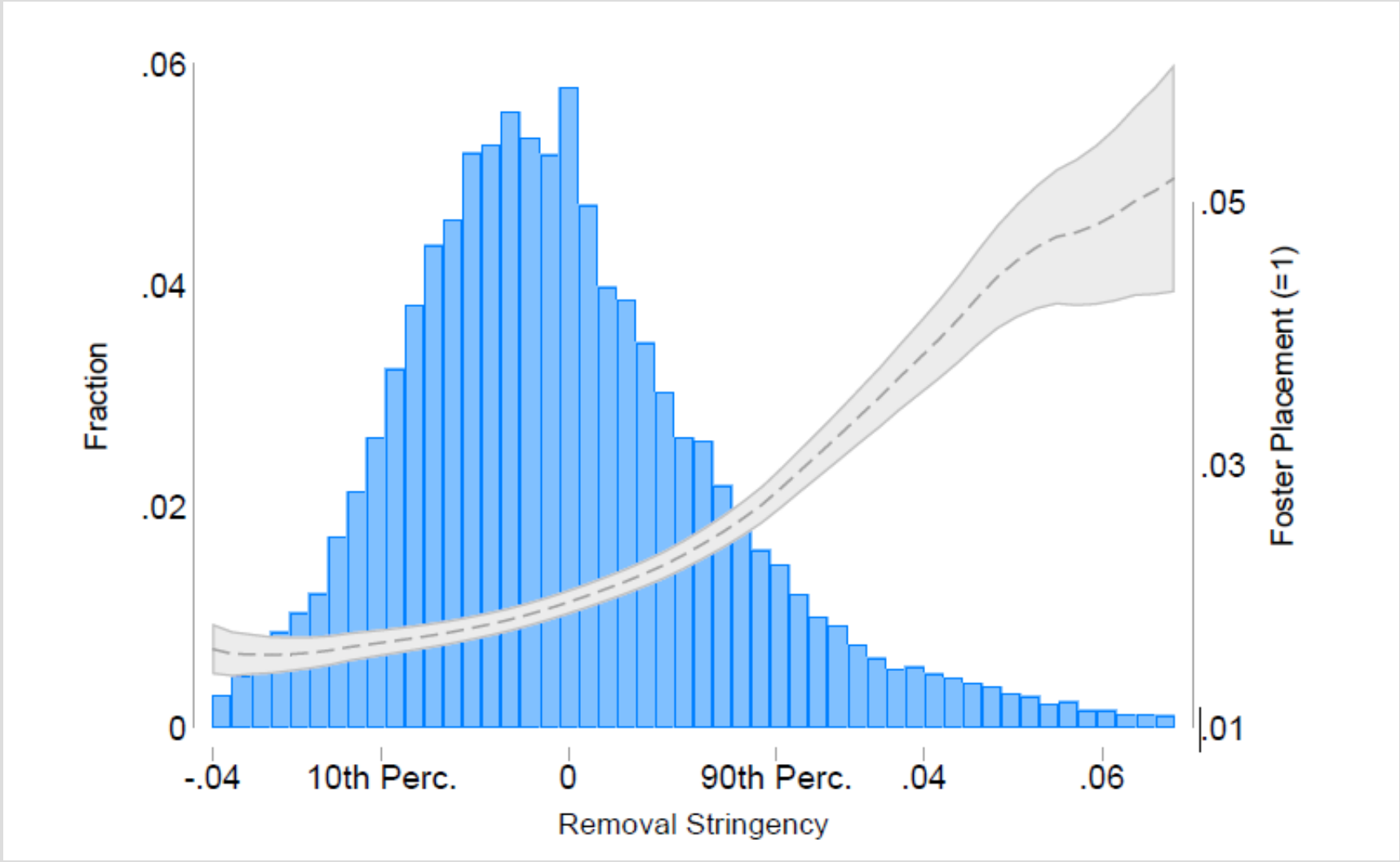
- Link to paper here: https://www.nber.org/papers/w29922
- Use administrative data (Michigan) to investigate foster care placement on likelihood of arrest and incarceration in adulthood

- Investigators (who are mostly randomly assigned to a given case) vary in their leniency/likelihood of sending to foster care



the 90th percentile removes at a rate 2.3 percentage points greater. Relative to the average removal rate of 3%, this indicates that moving from the 10th to the 90th percentile represents an almost 150% increase in the likelihood of placement.

Figure 2: Distribution of Investigator Removal Stringency Instrument



Baron and Gross (2022), p. 12

Table 7: Effects of Foster Care on Adult Convictions by Gender, Age, and Race/Ethnicity

	(1) Male	(2) Female	(3) Ages 6 to 11	(4) Ages 12 to 16	(5) White	(6) Black	(7) Hispanic
Foster Care	-0.496***	-0.114	-0.446***	-0.095	-0.243*	-0.253*	-0.282
	(0.189)	(0.094)	(0.135)	(0.140)	(0.144)	(0.133)	(0.346)
	{0.593}	{0.189}	{0.530}	{0.138}	{0.283}	{0.424}	{0.358}
Observations	59,976	58,297	52,675	65,598	74,127	33,045	5,975

Notes. Each column reports estimates from a separate 2SLS regression of whether a child was convicted by age 19 on foster care. All regressions include zip code by investigation year fixed effects and the covariates listed in Table A10. Control complier means are in curly brackets. Standard errors are clustered by child.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Baron and Gross (2022), p. 18

USEFUL RESOURCES

- Cunningham, *Causal Inference: The Mixtape*
  - Online version: https://mixtape.scunning.com/
- Angrist and Pischke, *Mostly Harmless Econometrics*
  - Website: https://www.mostlyharmlesseconometrics.com/
- Huntington-Klein, *The Effect*
  - Online version: https://theeffectbook.net/
  - His YouTube videos are great too: https://www.youtube.com/@NickHuntingtonKlein

QUESTIONS?

GARRETT BAKER
PHD CANDIDATE, DUKE UNIVERSITY

GARRETT.BAKER@DUKE.EDU

NEXT WEEK…

July 26, 2023

Presenter:

Frank Edwards,
Rutgers University

Topic:

Missing Data in R